It's Written On Your Face: Detecting Affective States from Facial Expressions while Learning Computer Programming

Nigel Bosch[1], Yuxuan Chen[1], Sidney D'Mello[1,2]

Departments of Computer Science[1] and Psychology[2], University of Notre Dame
{pbosch1, ychen18, sdmello}@nd.edu

Abstract. We built detectors capable of automatically recognizing affective states of novice computer programmers from student-annotated videos of their faces recorded during an introductory programming tutoring session. We used the Computer Expression Recognition Toolbox (CERT) to track facial features based on the Facial Action Coding System, and machine learning techniques to build classification models. Confusion/Uncertainty and Frustration were distinguished from all other affective states in a student-independent fashion at levels above chance (Cohen's kappa = .22 and .23, respectively), but detection accuracies for Boredom, Flow/Engagement, and Neutral were lower (kappas = .04, .11, and .07). We discuss the differences between detection of spontaneous versus fixed (polled) judgments as well as the features used in the models.

1 Introduction

Learning computer programming is an early obstacle for students pursuing a computer science (CS) degree [1]. The difficult nature of computer programming and lack of prior knowledge of novice students can create a particularly frustrating and confusing learning experience. One of the strategies that can be adopted to help with the burden of effectively teaching a large number of novice students is the use of intelligent tutoring systems (ITSs). As has been seen in other domains like computer literacy [2], it is likely that incorporating awareness of student affect into a computer programming ITS would lead to increased proficiency, particularly for novice students. Of course, an affect-aware ITS can never respond to affect if it cannot detect affect. We demonstrate a method for detecting the affect of novice programming students in a computerized learning environment using videos of students' faces.

Related Work. Affect detection can be done using various types of data sources, such as interaction data, speech, and physiology [3]. Facial-feature based affect detection is attractive because there is a strong link between facial features and affective states [4], it is more independent of learning environment or content (compared to interaction features), and it does not require expensive hardware, as webcams are ubiquitous on laptops and mobile devices.

In previous research on affect detection from facial features, Kapoor et al. [5] used multimodal data channels including facial features from video to predict Frustration in

an automated learning companion. They were able to predict when a user would self-report Frustration with 79% accuracy (chance being 58%). Hoque et al. [6] used facial features and temporal information in videos to classify smiles as either frustrated or delighted. They were able to accurately distinguish between Frustrated and Delighted smiles correctly in 92% of cases. They also found differences between posed (acted) facial expressions and naturally induced facial expressions. Only 10% of Frustrated cases included a smile in acted data, whereas smiles were present in 90% of cases of naturally occurring Frustration.

The Computer Expression Recognition Toolbox (CERT) [7] is a computer vision tool used for automatic detection of 19 Action Units (AUs, codes describing specific facial muscle activations) as well as head pose and position information. It also supplies measures of three unilateral (one side of the face only) AUs, as well as "Fear Brow" and "Distress Brow," which indicate the presence of combinations of AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), and AU4 (Brow Lowerer). CERT has been tested with databases of both posed facial expressions and spontaneous facial expressions, achieving accuracies of 90.1% and 79.9% respectively when discriminating between video frames with the presence vs. absence of particular AUs.

Whitehill et al. [8] have used CERT to detect engagement in a learning session. They obtained fine-grained judgments of engagement from external observers and achieved an accuracy of 71.8% when classifying instances of engagement vs. no engagement using AUs detected by CERT. Additionally, they found a correlation (r = .42) between fine-grained difficulty self-reports from students and AUs detected by CERT.

Grafsgaard et al. [9] used CERT to detect the overall level of Frustration (self-reported on a Likert scale) present in a learning session with modest results (R^2 = .24). Additionally, they have achieved good agreement between the output of CERT AU recognition and human-coded ground truth measurements of AUs (Cohen's kappa >= .68 for several key AUs), thereby providing additional evidence of the validity of CERT for automated AU detection.

Current Approach. This paper differs from the previous work in that our detectors will be applied at a finer granularity (15-second intervals), recognizing instances of affective states within a learning session rather than the level reported for the entire session. Additionally, the affective states we track are predominately learning-centered rather than the more commonly detected basic emotions (e.g. anger, sadness). Confusion is especially unique in that to our knowledge automatic confusion detection has not previously been done at a fine-grained level using facial features.

The facial expressions in the present study are naturalistic expressions, which have been shown to be more difficult to detect than posed expressions [10]. Despite the difficulty, we propose that facial features can be an effective method of automatically distinguishing particular affective states others in the domain of computer programming. To explore this we will answer these research questions: 1) Which affective states can be detected? 2) Can detection be improved by considering the type of affect judgment that is made? 3) Which features are most useful for detecting affective states?

2 Method

Data Collection. The data was collected from 99 computer programming novices who used a computerized learning environment designed to teach the basic elements of computer programming in the Python language. After the learning session, students viewed synchronized videos of their face and on-screen activity that had been recorded during the learning session, and retrospectively self-reported affective states at fixed points in the learning session. These periods were chosen to correspond with interaction events, such as typing code or viewing a new exercise, as well as idle periods (periods of time with no interaction events for more than 15 seconds). Students were also allowed to make spontaneous affect judgments at any point in the retrospective affect judgment process if they chose to. This retrospective affect judgment protocol allows for judgments to be made on the basis of a combination of the students' facial expressions, contextual cues (via screen capture), and their memories of the learning session [11]. We found that five affective states, namely Boredom (9%), Confusion (21%), Flow/Engagement (24%), Frustration (12%), and Neutral (17%), formed 83% of the affective states reported, so we focused on detecting these states (see [12] for more comprehensive details on data collection methodology used in the current study).

Computing Facial Features. We used CERT to calculate occurrence likelihoods for AUs in each video frame. The output of CERT was z-standardized within students and temporally aligned with affect judgments, then divided into segments of variable length (see below), each leading up to each affect judgment. Features were calculated by aggregating frame-level AU likelihoods across each segment using the median, maximum, and standard deviation of the 19 AUs provided by CERT. Head orientation and nose position were included as well. We also used HAAR cascades to detect the size of the face and the visibility of nose, mouth, eyes, and ears to provide additional information for situations with unusual pose or occlusion. We eliminated features exhibiting multicollinearity (variance inflation factor > 5).

Supervised Classification. We used the segment-level aggregate features to build classification models with the Waikato Environment for Knowledge Analysis (WEKA), a popular machine learning tool. Leave several out student-level cross-validation was used for model validation, with data from 66% of students randomly chosen to train classifiers and the remaining data used to test the performance of the classifiers. This ensures that the models generalize to new students since training and testing data sets are independent. The models were each trained and tested over 50 iterations with random students chosen each time to amortize random sampling error.

RELIEF-F feature ranking was used on the training data for each of the 50 iterations in order to identify the most diagnostic features prior to classification. We used 15 different classifiers and 6 different video segment sizes (2, 3, 6, 9, 12, and 15 seconds) to determine which segment size was likely to work best for a particular classification task. We then attempted to improve each of the best data configurations by either oversampling the training data (with SMOTE [13]) or downsampling the training data to equal class proportions.

3 Results and Discussion

Question 1: Which affective states can be detected? We attempted to individually classify each of the five most common affective states compared to all other affective states combined ("Other", which includes rare affective states). Cohen's kappa was used as the primary measure of performance, because it is more robust to class imbalances. Kappa measures the agreement between predicted and actual labels compared to chance (kappa = 0), with 1 reflecting perfect detection.

Classification of affective states from fixed affect ratings was not very successful. Using only fixed affect judgments, we had 6000 instances to be split into training and testing sets. Flow/Engagement was classified best (kappa = .112), with Boredom, Confusion, Frustration, and Neutral (kappas = .038, .064, .083, .070) classifications barely above chance. Classification of these data was expected to be difficult because they were selected at fixed points that were mostly independent of any facial activity. To improve the efficacy of classifiers we built models made using the spontaneous affect judgments, as discussed next.

Question 2: Can detection be improved by considering the type of affect judgment that is made? Because spontaneous affect judgments come from points in time of the student's choice, they may represent noticeable facial features in the video streams, as previously documented [14]. These judgments would likely make a more viable task for facial-feature based affect detection than the fixed judgments.

Only Confusion and Frustration had at least 100 spontaneous affect ratings, so we only consider those two states for further analysis. For the "Other" affective states, we sampled randomly from the fixed affective state judgments (5 times for each of the 50 iterations). Spontaneous Confusion judgments thus composed 21% of 582 instances with the other affective states in the fixed distribution as well, while Frustration composed 12% of 527 instances.

Using spontaneous judgments in this manner we were able to detect Confusion and Frustration much more effectively (kappa = .221 and .232, respectively). A simple logistic classifier yielded the best model for detecting Confusion, while an updatable naïve Bayes classifier was the most effective for Frustration. The best segment size for both of these was short (2 seconds for Confusion, 3 seconds for Frustration). Feature selection was used to select 50% of features for Confusion and the best 25% for Frustration detection. A more detailed look at the performance of these classification models can be found by examining the confusion matrices in Table 1.

Table 1. Confusion matrix for Confusion and Frustration spontaneous judgments

	Predicted Confusion	**Predicted Other**	**Priors**
Actual Confusion	0.50 (hit)	0.50 (miss)	0.21
Actual Other	0.25 (false alarm)	0.75 (correct rejection)	0.79
	Predicted Frustration	**Predicted Other**	
Actual Frustration	0.40 (hit)	0.60 (miss)	0.12
Actual Other	0.13 (false alarm)	0.87 (correct rejection)	0.88

Both models were impressive in terms of the low false alarm rate (i.e., Other affective states incorrectly detected as Confusion or Frustration). These models accurately detected half of the Confusion instances and nearly half of the Frustration instances, while properly rejecting most of the Other affective states, despite class imbalances.

Question 3: Which features were most useful for detecting affective states? We examined the features that were automatically selected for the spontaneous affect judgment classifiers. While both classification tasks used AU45 (Blink) features frequently, the other features differed between Confusion and Frustration. Particularly notable were the presence of unilateral (one side of the face) features for Frustration detection as well as head pose features (Yaw). Distress Brow, which appears frequently for Confusion detection, indicates evidence for AU1 (Inner brow raiser) or a combination of AU1 and AU4 (Brow lowerer). This feature has been found to be predictive in prior research involving manual coding of AUs as well [4]. Additionally, it appears as though Confusion was manifested more in absolute values of facial features, while Frustration was more easily detected from standard deviation features.

4 General Discussion

Despite the complexities of affect detection of naturally occurring learning-centered states, we were able to achieve some success in building fully-automated facial-feature based detectors of spontaneous confusion and frustration in a manner that generalizes to new students. Our current detection accuracy is modest at best, but affect detection is inherently an imperfect science and current detection rates are comparable with what is achieved for automatic detection of naturalistic affect from alternate modalities in a student-independent fashion [10].

Results for spontaneous judgments were much improved over the fixed judgments, as was expected. A similar phenomenon occurs when examining the inter-rater reliability between human judges manually coding emotions or AUs in video [14]. These results suggest that future work collecting video data for building affect detectors might be better served by focusing only on spontaneous affect judgments.

Some limitations of this study include (1) the relative infrequency of spontaneous judgments, (2) the relatively small sample size, and (3) lack of generalizability of results beyond the current sample. In future work we plan to train detectors using interaction data from the students' learning sessions, and incorporate those detectors with video-based detectors to create a more powerful multimodal affect classifier. We hope that accurate affect detection for novice computer programmers will lead to more effective computerized learning environments capable of responding to the momentary affective episodes of students so they may learn to their fullest potential.

Acknowledgment. This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958) and the Bill & Melinda Gates Foundation. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

References

1. Rodrigo, M.M.T., Baker, R.S.J. d, Jadud, M.C., Amarra, A.C.M., Dy, T., Espejo-Lahoz, M.B.V., Lim, S.A.L., Pascua, S.A.M.S., Sugay, J.O., Tabanao, E.S.: Affective and behavioral predictors of novice programmer achievement. SIGCSE Bulletin. 41, 156–160 (2009).
2. D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., Graesser, A.: A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In: Aleven, V., Kay, J., and Mostow, J. (eds.) Intelligent Tutoring Systems. pp. 245–254. Springer, Berlin Heidelberg (2010).
3. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence. 31, 39–58 (2009).
4. McDaniel, B.T., D'Mello, S.K., King, B.G., Chipman, P., Tapp, K., Graesser, A.C.: Facial features for affective state detection in learning environments. Proceedings of the 29th Annual Cognitive Science Society. pp. 467–472 (2007).
5. Kapoor, A., Burleson, W., Picard, R.W.: Automatic prediction of frustration. International Journal of Human-Computer Studies. 65, 724–736 (2007).
6. Hoque, M.E., McDuff, D.J., Picard, R.W.: Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles. IEEE Transactions on Affective Computing. 3, 323–334 (2012).
7. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The computer expression recognition toolbox (CERT). 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011). pp. 298–305 (2011).
8. Whitehill, J.R.: A stochastic optimal control perspective on affect-sensitive teaching. PhD dissertation, University of California, San Diego (2012).
9. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Automatically Recognizing Facial Indicators of Frustration: A Learning-Centric Analysis. (2013).
10. D'Mello, S., Kory, J.: Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. Proceedings of the 14th ACM international conference on Multimodal interaction. pp. 31–38. ACM, New York, NY, USA (2012).
11. D'Mello, S., Graesser, A., Picard, R.W.: Toward an affect-sensitive AutoTutor. Intelligent Systems, IEEE. 22, 53–61 (2007).
12. Bosch, N., D'Mello, S., Mills, C.: What Emotions Do Novices Experience during Their First Computer Programming Learning Session? In: Lane, H.C., Yacef, K., Mostow, J., and Pavlik, P. (eds.) Artificial Intelligence in Education. pp. 11–20. Springer, Berlin Heidelberg (2013).
13. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16, 321–357 (2011).
14. Graesser, A.C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., Gholson, B.: Detection of emotions during learning with AutoTutor. Proceedings of the 28th Annual Meetings of the Cognitive Science Society. pp. 285–290 (2006).